COVER SHEET

A SPATIAL SIMILARITY RANKING FRAMEWORK FOR SPATIAL METADATA RETRIEVAL

A.R. Walker[A]*, M. Moody*[A] *and B. Pham*[A]
[A] Queensland University of Technology, Brisbane, Australia.

Abstract
Recently there has been a trend to facilitate access to large government repositories of spatial data[1]. In addition, there is an increasing number of non-government spatial providers who provide access to large quantities of spatial data for little or no cost. These factors have promoted research into improving metadata standards and metadata retrieval.

Unfortunately, current efforts have mainly focused on improving the textual description and subsequent textual query matching retrieval frameworks for spatial metadata. Consequently, current metadata standards only include spatial extent and spatial reference information. This has resulted in the current standards having limited spatial querying functionality.

Spatial data by definition is spatial aware, therefore, the spatial component should be exploited as much as possible to allow more complex spatial querying of spatial metadata. This paper presents an extension to the existing ISO 19115 metadata schema for Geographic Information and details a framework that ranks the spatial similarity between a query and spatial metadata. The framework utilises an *object frequency* (*of*) and *inverse spatial frequency* (*isf*) which are incorporated into the spatial metadata schema. This novel methodology is based on the *tf-idf* method used for text-based information retrieval. A Bayesian inference retrieval engine is utilised to rank the similarity between spatial query and metadata.

This contribution will allow complex spatial queries of spatial metadata. The spatial metadata will be ranked by spatial similarity which will improve the performance and efficiency of the spatial retrieval process.

Introduction
Spatial metadata retrieval techniques have not kept pace with the explosion of cheap and freely available spatial data on the Internet. Current metadata retrieval techniques have focused on improving the textual metadata description and subsequent textual key term retrieval frameworks. Consequently, the current retrieval techniques only include limited spatial query mechanisms and simplistic spatial relevance evaluation procedures. Typically, spatial queries are based on a match between a bounding box representation of the area covered by a dataset (spatial extent), and a rectangular query area (Schlieder and Vögele 2002). All datasets for which the corresponding bounding box intersects with the query area are assumed to be spatially relevant with respect to the query. With the number of available datasets rapidly growing, more complex spatial queries and improved spatial relevance procedures are required to facilitate efficient retrieval of spatial data.

This paper presents a spatial metadata retrieval framework. The novelty of this research is the integration of additional spatial information into the metadata standard and the use of an original *of-isf* ranking strategy. The significance of this work is its ability to allow complex spatial querying of metadata which will improve the performance and efficiency of the spatial metadata retrieval process.

The remainder of this paper is structured as follows. Firstly, constraints of current metadata standards and current spatial metadata retrieval frameworks are reviewed. Then, the Materials and Methods section outlines the proposed spatial metadata retrieval framework and the experiment conducted to evaluate it. This is followed by a discussion of the results of the research. Finally, the conclusions and future work are presented.

Background
This section will investigate current metadata standards and retrieval methods. It will also investigate the advantages of using the spatial component in metadata searches.

[1] For example, the Australian Spatial Data Infrastructure framework and ANZLIC metadata project aim to improve access to Australia's spatial repositories.

Metadata

Metadata is data about data. It is information that describes the content, quality, condition, origin, and other characteristics of data or other pieces of information. It provides a way to document a dataset so that potential users of that data will be able to find it and evaluate it to see if it is suitable for their needs. Metadata is one of the main components in any retrieval framework and facilitates sharing data and knowledge locally, across networks or across the Internet.

Current Metadata Standards

The three main spatial metadata standards that were considered in this research were developed by ANZLIC, FGDC and ISO respectively. Each metadata standard is explained in the following paragraphs.

The Australia New Zealand Information Council (ANZLIC) developed and maintains the ANZLIC Metadata Guidelines for the Australian and New Zealand governments (ANZLIC 2001). These guidelines were developed to define the minimum requirements for metadata to be included in the Australian Spatial Data Directory (ASDD). Recently ANZLIC has started expanding its guidelines to include the broader International Standards Organisation (ISO) 19115 standard. This will involve defining an Australian and New Zealand profile of ISO 19115. This profile is called the ANZLIC ISO Metadata Profile and a draft version was issued earlier this year (ANZLIC 2005).

The Federal Geographic Data Committee (FGDC) developed the Content Standard for Digital Geospatial Metadata (CSDGM) as the principal metadata standard in the USA (FGDC 1998). This standard provides a complete description of a data source and is mandatory for Federal agencies and recommended for state and local governments in the USA. Because the USA had to deal with a large variation across states this standard is very complex and has proven difficult to implement in its entirety. As a result, various states and regions have created their own metadata standards to try to simplify the information that should be recorded (Green and Bossomaier 2002).

The International Standards Organisation (ISO) developed the ISO 19115, Geographic Information—Metadata standard (ISO 2003). This standard attempts to satisfy the requirements of all existing metadata standards. It allows for either general or detailed descriptions of data sources, it makes some allowances for describing resources other than data, and has a small number of mandatory elements. While the 19115 standard is finalised, work on a new ISO standard has already begun. The new standard, 19139, Geographic Information—Metadata—Implementation will integrate content from several different ISO standards and provide new specifications for the format in which ISO metadata is stored.

ISO 19115 was chosen as the standard used in this research because its development was based on the collective knowledge of all existing successful metadata standards that were being used to describe spatial data. For example it has combined the best parts of ANZLIC and FGDC into one standard. In addition, ISO 19115 has taken a more inclusive approach to the parameters now routinely used by the range of data providers offering an extensive and, as a result much more complex set of metadata definitions. The new standard includes all elements that any typical jurisdiction would use, either to describe the resource or in providing the descriptions to users. The international acceptance of ISO 19115 has been good and that is why it is the standard chosen for this research.

Limitations of Current Metadata Retrieval Frameworks

The ASDD provides an advanced mechanism for metadata retrieval for Australian data on its web site (ANZLIC 2005). This retrieval interface is typical of the numerous GIS clearinghouse services available on the Internet. However, as noted in Figure 1 below, the spatial query is limited to a simple single geographic extent with the spatial options of “overlaps any part of”, “is entirely within” and “completely covers”. In addition, no ranking of the spatial similarity between datasets and query is shown in the search results (see Figure 2).

The main reason only simple spatial querying is provided by the major GIS clearinghouse has to do with the technology required to support more complex spatial querying. Handling shape criteria usually requires a spatial object type in the database, and complex spatial query processing. However, point and rectangle queries can be handled as simple arithmetic processes on standard numeric fields (i.e. the North, South, East and West fields of Figure 1). Therefore, if only simple spatial queries are allowed, then standard database services can be used to deliver them (Plewe 1997).

Figure 1, ASDD Advance Metadata Search

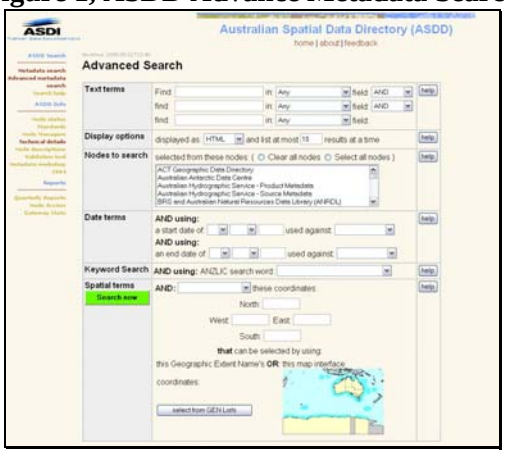


Figure 2, ASDD Search Results

Spatial catalogue retrieval frameworks are used to search metadata documents and assist potential data users in finding datasets that will best suit their needs. The metadata about the spatial datasets is stored in a dataset. Typically, metadata retrieval frameworks utilise standard database services such as an SQL-enabled web server or a Z39.50 server to query the metadata catalogue. For example, both FGDC's National Spatial Data Infrastructure Clearinghouse Network (FGDC 1995) and ANZLIC's ASDD use Z39.50. Z39.50 was designed for remote catalogue searches and has successfully been adopted to provide simple spatial metadata searches (NISO 2002). The main limitation of SQL-enabled or Z39.50 servers is that by using a standard database service only very simple spatial queries can be processed. Another example of spatial retrieval that focuses on matching the spatial query with only the spatial extent contained in the metadata document can be found in (Larson and Frontiera 2004).

The obvious solution would be to utilise spatially enabled database servers like ESRI's spatial database engine or Oracle's spatial database, however, the time to process complex spatial queries using accurate spatial objects is considerable and would require querying the whole dataset and not just the metadata. What is required is a process that can deliver a more complex spatial query but still utilise a simple arithmetic process to resolve said queries. This paper proposes one such solution. It will be shown that by adding some simple text elements to ISO 19115, a more flexible spatial XML metadata schema is obtained. This allows the proposed framework to extent beyond only considering the dataset's total spatial extent, but also including evaluating the similarity of spatial terms contained within the query and the dataset. This allows more complex queries and should lead to improved retrieval performance results.

Before moving onto the proposed framework, an explanation of a well known text retrieval ranking strategy is given. This strategy is important as it provides the basis of the new spatial retrieval strategy proposed in this paper.

tf-idf Ranking Strategy

The *tf-idf* (term frequency–inverse document frequency) ranking strategy is often used in text information retrieval (IR) systems (Salton and Buckley 1988) and is best explained through text searches on documents. This strategy is a statistical measure used to evaluate how important a word is to a document. The importance increases proportionally to the number of times a word appears in the document but is offset by how common the word is in all of the documents in the collection. *tf–idf* is often used by text search engines to find the most relevant documents to a user's text query (Baeza-Yates and Ribeiro-Neto 1999).

The *tf-idf* is characterised by two components. The first component is the term frequency or *tf* factor. It is a measure of the frequency of the term in the document and is calculated by,

$$tf_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \qquad \textbf{(1)}$$

where $freq_{i,i}$ is the raw frequency of the term k_i in document d_j. $tf_{i,j}$ is the normalised frequency of term k_i in document d_j (i.e. the number of times the term k_i is mentioned in the text document d_j). The $\max_l freq_{l,i}$ is calculated over all terms which are mentioned in the document. Because terms which appear in many documents are not as useful in distinguishing relevance, a second component for the method was introduced. It is the inverse document frequency or *idf* factor and is given by,

$$idf_i = \log \frac{N}{n_i} \tag{2}$$

where N is the total number of documents and n_i is the number of documents which contain the k_i term. Combining the *tf* and *idf* components we get the *tf-idf* term weight scheme which is given by,

$$tfidf_{i,j} = tf_{i,j} \times \log \frac{N}{n_i} \tag{3}$$

A high value in *tfidf* is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents, consequently, the strategy tends to filter out common terms. This paper presents a modified *tf-idf* ranking strategy that can be used for spatial metadata retrieval.

Materials and Methods

The main idea behind *tf-idf* is to match "key terms" or words in the document with "key terms" in the query. In a spatial sense, if the spatial domain can be defined as "spatial terms" then a similar method can be used to match "spatial terms" in GIS datasets with "spatial terms" in the query. By defining spatial terms in the metadata schema it will be shown that the existing Z39.50 or SQL servers can be used to resolve query to metadata matching criteria.

Spatial Terms

To enable simple arithmetic calculation of spatial queries a simple language of spatial terms was developed by splitting the world coordinates into 0.2° square cells. Each cell covers an area 20km x 20km or 400km². Each cell represents a unique spatial term denoted by its latitude/longitude position at the top left corner of the cell. The spatial terms that cover Australia and the Gold Coast are shown in Figure 3 and Figure 4 respectively.

Figure 3, Cells over Australia

Figure 4, Cells over Gold Coast

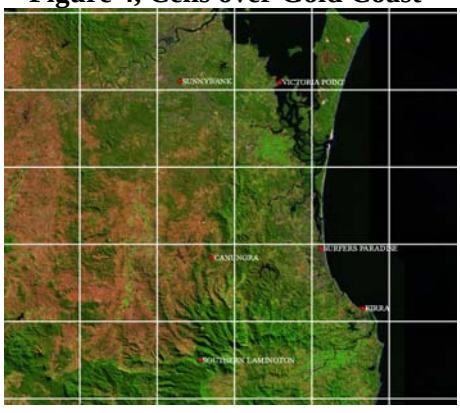
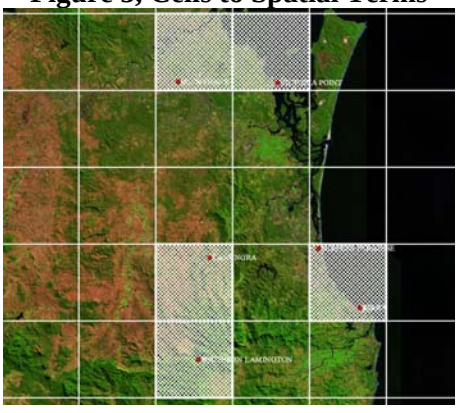


Figure 5, Cells to Spatial Terms



The cell size has been chosen to make the spatial term language's vocabulary similar in size to the English language. When using the *tf-idf* approach for text IR, the English language has approximately 2,000,000 words which are potential keys terms. The spatial term language presented here has 1,620,000 potential spatial terms. To illustrate how the spatial term language can describe a dataset, consider the *place names* dataset shown in Figure 4 which consists of six place names (i.e. Sunnybank, Victoria Point, etc). From Figure 4, it is obvious that five spatial terms, as shown in Figure 5, can be used to define the *place names* dataset.

With a suitable simple spatial term language defined, the next stage was to define a new ranking strategy for this simple spatial language. This new spatial ranking strategy will be referred to as *of-isf* throughout this paper and is explained in the next section.

of-isf Ranking Strategy
The *of-isf* (object frequency-inverse spatial frequency) is a novel way of adapting *tf-idf* to a simple spatial language. Defining the values of object frequency (*of*) and inverse spatial frequency (*isf*) parameters is the main distinction between this method and the *tf-idf* approach.

In the *tf*-idf approach the word frequency of a key term appearing in a document is used as a similar measure. Unlike words, spatial terms can only appear once to describe a spatial dataset as shown in Figure 5. In order to distinguish between GIS datasets that are described by exactly to same spatial terms, the frequency of the objects within each spatial term is used as a similarity measure.

Consequently, in the *of-isf* the approach the spatial importance is related to the frequency of the objects contained by each spatial term. The *of* factor is calculated by,

$$of_{i,j} = \frac{objfreq_{i,j}}{\max_l objfreq_{l,j}} \quad \textbf{(4)}$$

Although the equations for *tf* and *of* are basically the same, the calculation of their parameters is very different. Here $objfreq_{i,i}$ is the raw frequency of objects contained within spatial term s_i in dataset d_j. $of_{i,j}$ is the normalised frequency of objects contained within spatial term s_i in dataset d_j. The $\max_l objfreq_{l,i}$ is calculated over all spatial terms which are mentioned in the dataset. Again because spatial terms will appear in many datasets, and inverse spatial frequency or *isf* factor is given by,

$$isf_i = \log\frac{N}{n_i} \quad \textbf{(5)}$$

where N is the total number of datasets and n_i is the number of datasets which contain the s_i spatial term. Combining the *of* and *isf* components we get the *of-isf* term weight scheme which is given by,

$$ofisf_{i,j} = tf_{i,j} \times isf_i \quad \textbf{(6)}$$

A high value in *ofisf* is reached by a high object frequency (in the given spatial term within a given dataset) and a low spatial term frequency within the whole collection of datasets. With the ranking strategy finalised an IR methodology was chosen from the three main types of Boolean, vector and probabilistic (Korfhage 1997). Both vector and probabilistic IR methods are well suited to *tf-idf* ranking strategies (Baeza-Yates and Ribeiro-Neto 1999). Eventually, a Bayesian inference probabilistic IR method was chosen as it handles uncertainty or missing data better than the vector model.

Bayesian Inference Retrieval Model
Bayesian inference models the retrieval process as an evidential reasoning process (Turtle and Croft 1990; Turtle and Croft 1991). It associates random variables with the spatial terms, datasets and user queries. The datasets are observed individually as evidence and the degree of belief in the query is calculated and ranked for each dataset (i.e. calculate $P(q|d_j)$). $P(q|d_j)$ is the probability of the dataset given that the query has been observed. The datasets that return the highest degree of belief in the query are the datasets that are retrieved by the system as the relevant datasets for that query. More details on the Bayesian inference model used will be given later, but now the overall algorithm for spatial metadata retrieval is presented.

The Algorithm for Spatial Metadata Retrieval using Bayesian Inference
The algorithm for ranking the datasets using Bayesian Inference with an *of-isf* ranking strategy is as follows:

1) Convert the input query, q, into its spatial terms, s_j
2) For each dataset, d_j, in the metadata catalogue
    a. Calculate the spatial terms, s_j, it has in common with the query, q
    b. Build the Bayesian network
    c. Calculate the a priori and conditional probability tables
    d. Calculate $P(q|d_j)$
3) Rank $P(q|d_j)$

Main Components of Framework

The remainder of the Material and Methods section will explain in detail the main components of the proposed spatial metadata retrieval framework. The main components are:

1. GIS Data
2. GIS Metadata Catalogue
3. Bayesian Inference Retrieval Model
4. Metadata *of-isf* Ranking Strategy
5. Object Frequency Schema
6. Inverse Spatial Frequency Schema
7. Spatial Query
8. Ranked Query Output

GIS Data

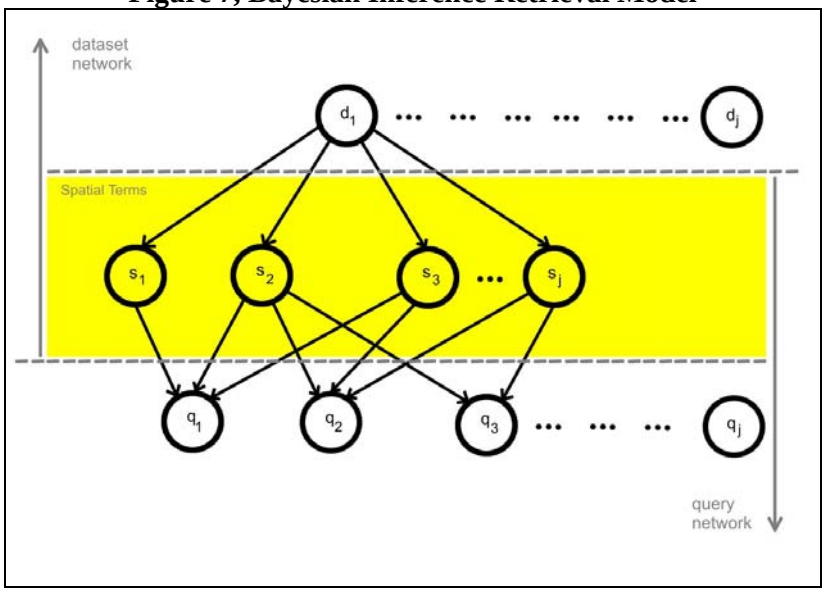
The GIS data component refers to a collection of distributed GIS data on the Internet. The GIS data can be either vector or raster format. Each GIS dataset must have an XML metadata document in ISO 19115 format describing it. The datasets themselves may be distributed across various location and organisation. This experiment used a small set of GIS data obtained from Gold Coast City Council.

Metadata Catalogue

The metadata catalogue is a registry of all the datasets available to the spatial metadata retrieval system. Only GIS datasets registered in the catalogue can be retrieved. The metadata catalogue contains all the metadata for all the GIS datasets in ISO 19115 format. Currently the ISO 19115 format only specifies the spatial extent of the dataset and the spatial reference system used. A typical GIS metadata XML document is shown in Figure 6. The advantage of XML is that it is extensible. This is why it was adopted as the metadata language in the first place. Before we can look at the proposed extensions to the metadata standard we must understand how the spatial retrieval will be achieved using a Bayesian Inference Retrieval Model.

Figure 6, Typical XML Metadata Document

Figure 7, Bayesian Inference Retrieval Model



Bayesian Inference Retrieval Model

The Bayesian inference retrieval model presented in this paper is loosely based on work by Turtle and Croft (Turtle and Croft 1990; Turtle and Croft 1991) Turtle and Croft's work was only intended for text document retrieval and did not contain any provision for retrieving spatial data. This research has adapted Turtle and Croft's method to include spatial evidence in the inference model. This is the first time that spatial relationships have been included into an inference model of this type.

The proposed Bayesian inference retrieval model is shown in Figure 7. This model utilises only spatial parameters to determine query to dataset similarity. The model associates random variables with datasets, spatial terms, and user queries. The dataset is the root node of the network. The Bayesian network is broken into two sub-network called dataset network and query network. The dataset network is constructed from the dataset's metadata. Each dataset is made of spatial terms, and has a causal relationship with them. An arc from the dataset to a spatial term indicates that there is a causal relationship between that dataset and that spatial term and that the observation of one causes a change in belief of the other. A random variable associated with dataset d_j is the probability of observing that dataset. The observation of the dataset d_j asserts a belief upon the random variable associated with its spatial terms. The observation of a dataset causes an increased belief in its spatial terms.

In a similar fashion the query network is constructed with arcs from spatial terms to query nodes for each of the spatial terms that have a relationship to particular queries. To simplify the calculations, it is assumed that all the random variables are binary.

In order to rank the similarity of a query to a dataset, a Bayesian network is created for each dataset-query pair. Then the probability of the dataset given that the query has been observed, $P(q|d_j)$, is calculated for each dataset. These probabilities are ranked to reveal which datasets best match a particular query.

The main difference between Turtle and Croft's work is the calculation of the a priori and conditional probabilities within the Bayesian network. The method use to calculate the a priori and conditional probability parameters is detailed in the next few paragraphs.

Because the datasets are at the root of the network, we only need to calculate the a priori probability for them. If we have no prior knowledge or preference about datasets, then the a priori probability is normally considered uniform. Thus,

$$P(d_j) = \frac{1}{N}$$
$$P(\bar{d}_j) = 1-\frac{1}{N} \quad \textbf{(7)}$$

where N is the total number of datasets in the metadata catalogue. The spatial term nodes require conditional probabilities to be calculated. This would normally require $O(2^n)$ space with a node with n parents. However, if the Noisy-OR method is used this is reduced to O(n) (Russell and Norvig 2003). As outlined in (Baeza-Yates and Ribeiro-Neto 1999), different calculation of the conditional probabilities can make the Inference network subsume the Boolean or Vector models. The following equations calculate the conditional probabilities based on the vector model and using the *of-isf* strategy. They are,

$$P(s_i|d_j) = of_{i,j}$$
$$P(\bar{s}_i|d_j) = 1- P(s_i|d_j) \quad \textbf{(8)}$$

$$P(q|s_i) = isf_i$$
$$P(\bar{q}|s_i) = 1- P(q|s_i) \quad \textbf{(9)}$$

where $of_{i,j}$ is the object frequency and isf_i is the inverse spatial frequency as previously given in equations (4) and (5).

Metadata of-isf Ranking Strategy
As mentioned previously, the main difference between *tf-idf* and *of-isf* is in the calculation of the a priori and conditional probabilities within the Bayesian network. In order for this information to be available to the Bayesian inference model it must first be incorporated into the metadata XML schema. The next sections explain this process and schema.

Object Frequency Schema
A simple spatial term language is used to describe the spatial makeup of the data. The object frequency within each spatial term is calculated and stored in the metadata for each particular spatial dataset. An efficient method for recording this information was developed and required the following XML elements to be added to the ISO standard.

1) "object frequency" – contains the latitude/longitude pairs that represents spatial terms;
2) latitude – contains a list a longitudes that represents a spatial term and the number of records contained within that spatial term. XML format of this element is "N|S [digit][digit][digit].digit", and;
3) longitude – contains number of records within the spatial term represented by the latitude/longitude pair. XML format of this element is "E|W [digit][digit][digit].digit"

The XML element <objectfrequency> is the main node. As mentioned above it is broken down into latitude/longitude coordinate pairs that describe the 1,620,000 spatial terms on the earth's surface. The latitude is described in 0.2° increments for north (N0.0°, N0.2°, .. , N90.0°) and south (S0.0°, S0.2°, .. , S90.0°). The longitude is described in 0.2° increments for east (E0.0°, E0.2°, .. , E180.0°) and west (W0.0°, W0.2°, .. , W180.0°).

An example of the new XML schema and the spatial objects it represents is shown in Figure 8. From Figure 8, one can see that the dataset overlaps 12 spatial terms, however only 3 of these spatial terms contain objects. Therefore, only three spatial terms are required to describe the object frequency of the example dataset.

An example of the schema has been given using point objects. The schema can describe line and polygon objects as well. Both lines and polygons are described by the schema in terms of an overlap spatial relationship between the object and the spatial term. If a single object is in more than one spatial term it is added to each spatial term that overlaps it as shown in Figure 9 and Figure 10 for line and polygon objects respectively.

Figure 8, Point Object Frequency

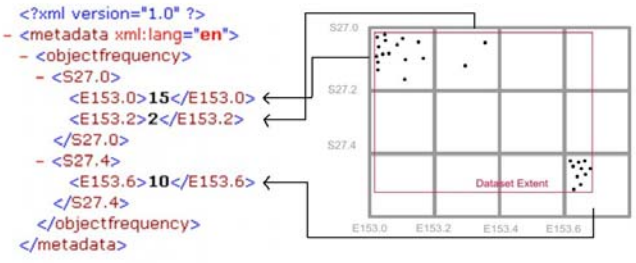
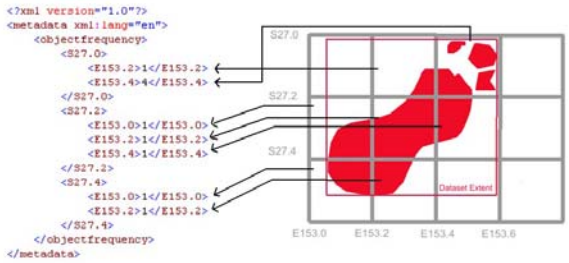


Figure 9, Polygon Object Frequency



As metadata is added to the metadata catalogue, the object frequencies are automatically added to the XML document. A program was written in MapObjects that calculated dataset's object frequencies for each of the 1,620,000 spatial terms. The algorithm used was:

1) from dataset's extent calculate the initial spatial extent to test as a multiple of 0.2°
2) divide into four equal sized square cells
3) for each cell, keep splitting where objects exist until cell size equals 0.2° x 0.2°
4) count object frequency in 0.2° x 0.2° cell which represent a spatial term

Inverse Spatial Frequency Schema
As the metadata is added to the metadata catalogue, the inverse spatial frequencies are updated automatically. A metadata XML document that summaries the metadata catalogue will contain the number of datasets in the catalogue

and inverse spatial frequencies will have the XML schema format shown in Figure 11. The "numberofdatasets" element is the total number of datasets contained in the metadata catalogue. The "inversespatialfrequency" element is the number of datasets in the metadata catalogue that have objects in that spatial term.

Figure 10, Line Object Frequency

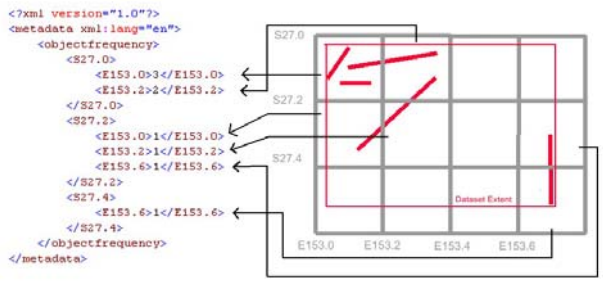
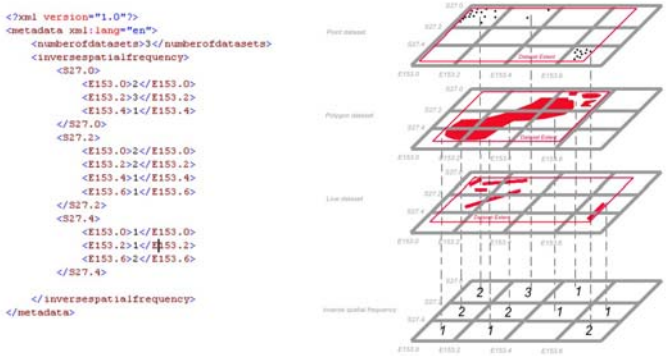


Figure 11, Inverse Spatial Frequency



Spatial Query

As mentioned previously, the current metadata retrieval frameworks only allow simple spatial searches because only the spatial extent of datasets is stored in the spatial metadata. This framework allows more complex spatial queries that can combine multiple spatial query areas as shown in Figure 12.

Figure 12, Multiple Spatial Query

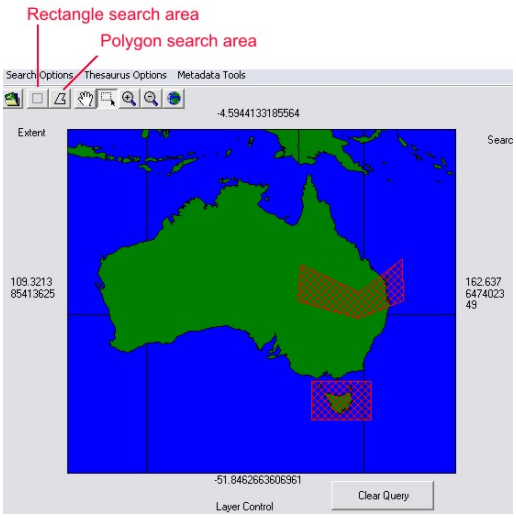
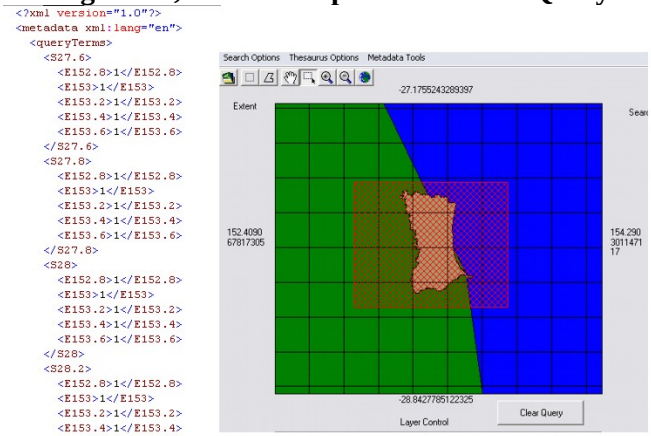


Figure 13, Calculate Spatial Terms of Query



The spatial query is broken into a set of spatial terms that describe the query as shown in Figure 13. The spatial query has the same format as the object frequency tags added to the dataset metadata file. An element value of "1" has been chosen but has no real significance other than to show the presence of spatial terms in the query.

Ranked Query Output

The Bayesian inference calculation of $P(q|d_j)$ will return a ranked list of datasets as shown in Figure 14. These datasets are listed with image thumbnail, dataset name, and other essential information about the spatial data. This format is similar to that used currently by the major GIS portals such of the Geography Network (ESRI 2005).

Prototype system

Finally all the components were combined into a prototype spatial metadata retrieval system. This prototype was developed in "c# .NET" utilising MapObjects (ESRI 2006) and MSBNx (Microsoft 2003).

Results and Discussion

A prototype spatial metadata retrieval system was developed. The user interface is shown in Figure 15 and includes spatial, textual and thesaurus search functionality. Only the spatial query functionality has been presented in this paper and both textual and thesaurus search methodologies will be presented in a future paper.

Figure 14. Ranked Retrieved List

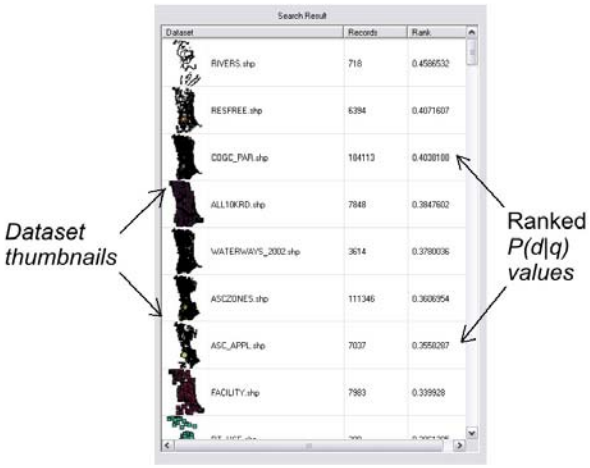
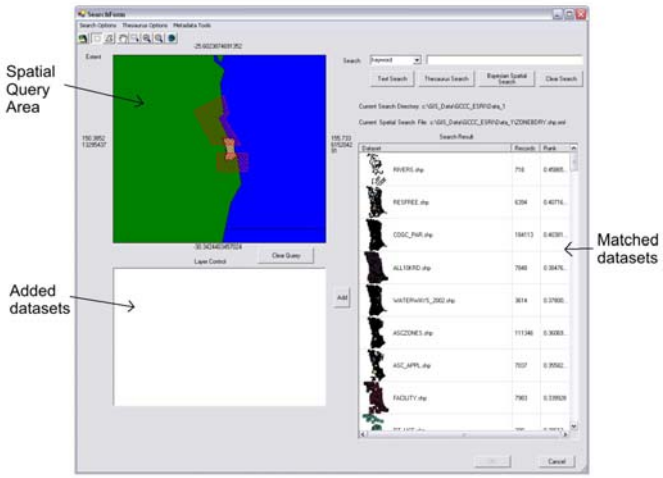


Figure 15. Prototype Spatial Retrieval System



From Figure 15, it can be seen that the user can enter a complex spatial area as the search query. In the spatial query area, *World continents* and *Gold Coast city council boundary* datasets have been added to provide a spatial reference to the user. The results of the query (including thumbnail) are shown in the matched datasets area of Figure 15. Finally, the user has the ability to add any retrieved dataset to the spatial query area map in order to assist the user via additional spatial reference information. In Figure 15, no additional datasets have been added and this area is shown as blank.

Initial results showed that the Bayesian inference and *of-isf* ranking strategy performed as expected. A test set of queries retrieved all anticipated datasets. For example, datasets with high object frequencies in the query's spatial terms and with corresponding low spatial term frequencies within the whole collection of datasets were ranked high in the retrieved results.

The overhead to the metadata document is small as only spatial terms that have objects contained within them are included. The advantage of *of-isf* is that object frequencies can be calculated in a pre-processing stage as the dataset is added to the metadata catalogue. Thus, the spatial metadata retrieval framework presented here only requires spatial information stored in a metadata file and does not process complex spatial queries on the underlying spatial data itself. This means that the queries run quickly and efficiently which is the major requirement in any IR system.

The cell size of the spatial terms used dictates the resolution of the spatial queries. This framework is not intended to replace standard spatial functions but is to serve as a technique to facilitate improved spatial querying using simple arithmetic servers. If the spatial cell size were reduced it would increase the resolution but would have an effect on the performance.

Conclusions

The spatial metadata retrieval framework presented in this paper will allow complex spatial queries by utilising the spatial component in the retrieval process. This research exploits the spatial component as much as possible with the existing SQL-enabled web and Z39.50 servers. It achieves this without the need to perform complex spatial calculations by using a 0.2º uniform grid to define a spatial term language for spatial metadata. In addition, the current ISO metadata schema was changed to incorporate these new spatial elements. The new elements had minimal effect on the overall performance of the non-spatial querying.

The novelty of this research is the unique spatial term language and *of-isf* ranking strategy that can easily be incorporated into the existing ISO 19115 standard. The significance of this contribution is its ability to allow complex spatial queries while still utilising the existing SQL-enabled web and Z39.50 servers. Finally, this proposed framework will improve the performance and efficiency of the spatial retrieval process by allowing users the find the most appropriate datasets for their information needs.

Future Work

Work has started to extend the current framework to include text key term similarity matching in a combined hierarchical retrieval system. This will include investigating the importance of text similarity versus the importance of spatial similarity in matching data to a query.

Acknowledgements

This research is supported by the Built Environment Research Unit of the Queensland Government. The Gold Coast City Council provided the GIS datasets for use in the experiment.

References

ANZLIC (2001). ANZLIC Metadata Guidelines: Core metadata elements for geographic data in Australia and New Zealand Version 2, http://www.anzlic.org.au/infrastructure_metadata.html, Accessed on 23 May 2006.

ANZLIC (2005). ANZLIC ISO Metadata Profile, http://www.anzlic.org.au/metadata_project.html, Accessed on 23 May 2006.

ANZLIC (2005). Australian Spatial Data Directory (ASDD), http://asdd.ga.gov.au/, Accessed on 15 June 2005.

Baeza-Yates, R. and B. Ribeiro-Neto (1999). *Modern Information Retrieval*. New York, ACM Press.

ESRI (2005). Geography Network, ESRI, http://www.geographynetwork.com/, Accessed on 31-May 2005.

ESRI (2006). MapObjects, Environmental Systems Research Institute, http://www.esri.com/, Accessed on 23 May 2006.

FGDC (1995). National Spatial Data Infrastructure Clearinghouse Network, http://www.fgdc.gov/dataandservices/, Accessed on 23 May 2006.

FGDC (1998). Base Content Standard for Digital Geospatial Metadata (version 2.0), FGDC-STD-001-1998, http://www.fgdc.gov/standards/standards_publications/, Accessed on 13 Mar 2006.

Green, D. and T. Bossomaier (2002). *Online GIS and Spatial Metadata*. London, Taylor & Francis.

ISO (2003). ISO 19115:2003 Geographic Information - Metadata, ISO Standards, http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35&ICS2=240&ICS3=70, Accessed on 13-Mar- 2006.

Korfhage, R. R. (1997). *Information Storage and Retrieval*. New York, John Wiley & Sons.

Larson, R. R. and P. Frontiera (2004). Geographic Information Retrieval (GIR) Ranking Methods for Digital Libraries. *4th ACM/IEEE-CS joint conference on Digital libraries, Tuscon, AZ, USA*, ACM Press.

Microsoft (2003). MSBNx: Bayesian Network Editor and Toolkit, Microsoft, http://research.microsoft.com/adapt/msbnx/default.aspx, Accessed on 1 Dec 2003.

NISO (2002). Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Maryland, NISO Press, http://www.loc.gov/z3950/agency/, Accessed on 23-May 2006.

Plewe, B. (1997). *GIS Online: Information Retrieval, Mapping, and the Internet*. New York, OnWord Press.

Russell, S. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach*. London, Prentice Hall.

Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal* **24**(5): 513-523.

Schlieder, C. and T. Vögele (2002). Indexing and Browsing Digital Maps with Intelligent Thumbnails. *International Symposium on Spatial Data Handling (SDH) 2002, Ottawa, Canada*, Springer.

Turtle, H. and W. B. Croft (1990). Inference networks for document retrieval. *13th annual international ACM SIGIR conference on Research and development in information retrieval, Brussels, Belgium*, ACM Press.

Turtle, H. and W. B. Croft (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems (TOIS)* **9**(3): 187 - 222.